## **Overview**

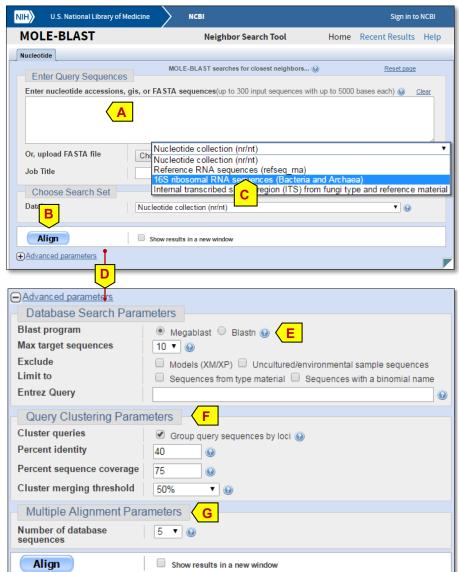
Large scale sequencing of unknown or environmental samples is a powerful method in biological research to help identify organisms present on the skin, in the gut, in soil sample, or in harsh environments such as toxic sites and natural ecosystems, to provide important insight on human health, bioremediation, global and local ecosystem processes. MOLE-BLAST is a tool designed cluster input sequences and place them in taxonomic context to help identify these sequences and the organisms they were isolated from. It works by taking multiple nucleotide sequences as input, searching them against a source database to find their best matches. The input query sequences along with their newly identified database matches are then subjected to multiple sequence alignment and the result is used to construct a phylogenetic tree. By placing input query sequences under the context of known sequences, MOLE-BLAST helps establish the taxonomic relationship to those known sequences and significantly increases the reliability of the identification process.

## **Access**

MOLE-BLAST is available from the "Specialized BLAST" section of the BLAST homepage (<u>blast.ncbi.nlm.nih.gov</u>). The search page is at <u>blast.ncbi.nlm.nih.gov/moleblast/moleblast.cgi</u>.

The top input box (A) of this page accepts a group of query nucleotide sequences in FASTA format or a list of NCBI database identifiers. Clicking the "Align" button (B) starts the search with default settings with the neighboring sequences coming from NT nucleotide database. MOLE-BLAST also provides specialized databases including the 16S reference sequences from bacteria and archaea (C). Additional parameters for the search are available in the expandable "Advanced parameters" section (D). The "Database Search Parameters" section (E) controls the search sensitivity - megablast for faster searches, blastn for more sensitive searches, and number of matches to keep for the BLAST search step. It also allows database subset selection for more focused search. The "Query Clustering Parameters" section (F) sets thresholds for separating clusters into loci. These settings may separate sequences from different genes if they are present in the sample. The "Multiple Alignment Parameters" section (G) determines how many database matches found in the BLAST search step should be used in multiple alignment to construct the phylogenetic tree.

Pushing the "Align" button initiates the search and brings up the interim screen (H). Leave this browser window open until the search is completed to avoid loosing the assigned Request ID (RID, I) and the results. Search results are save for 24 hours and can be retrieved using the assigned RID and the input box under the "Recent Results" tab.



Please, do not close the browser before the processing is completed

HC1DWV3J413

Fri Mar 27 21:36:55 2015

Fri Mar 27 21:38:45 2015

Contact: blast-help@ncbi.nlm.nih.gov

1 minute 50 seconds

Calculating

Mole-BLAST Request ID

Time since submission

Status

Submitted at

Current time

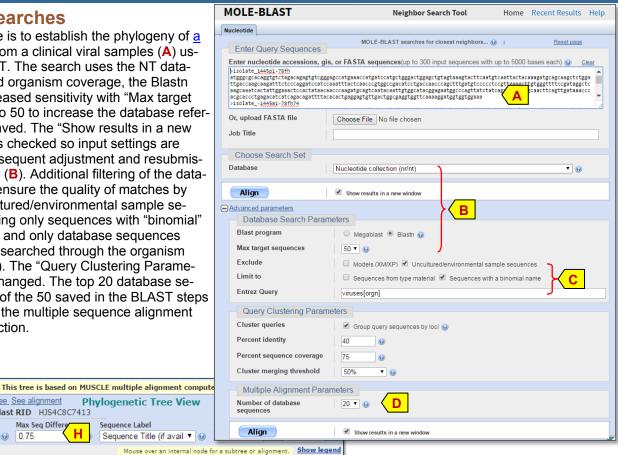
# **Example Searches**

Reset Tree Download Tree See alignment

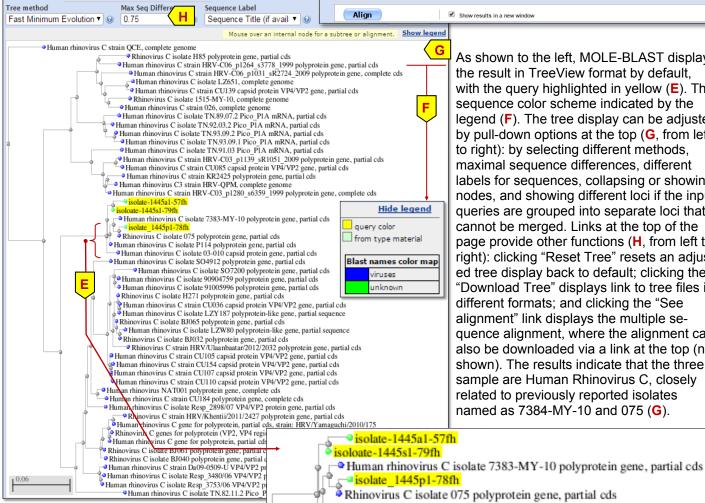
Moleblast RID HJS4C8C7413

Page 2

The first example is to establish the phylogeny of a set sequences from a clinical viral samples (A) using MOLE-BLAST. The search uses the NT database for its broad organism coverage, the Blastn program for increased sensitivity with "Max target sequences" set to 50 to increase the database reference matches saved. The "Show results in a new window" option is checked so input settings are available for subsequent adjustment and resubmission if necessary (B). Additional filtering of the database is done to ensure the quality of matches by excluding "Uncultured/environmental sample sequences," selecting only sequences with "binomial" scientific names, and only database sequences from viruses are searched through the organism Entrez Query (C). The "Query Clustering Parameters" are left unchanged. The top 20 database sequences (D) out of the 50 saved in the BLAST steps are requested in the multiple sequence alignment and tree construction.



**MOLE-BLAST** 



**Phylogenetic Tree View** 

As shown to the left, MOLE-BLAST displays the result in TreeView format by default, with the guery highlighted in yellow (E). The sequence color scheme indicated by the legend (F). The tree display can be adjusted by pull-down options at the top (G, from left to right): by selecting different methods, maximal sequence differences, different labels for sequences, collapsing or showing nodes, and showing different loci if the input queries are grouped into separate loci that cannot be merged. Links at the top of the page provide other functions (H, from left to right): clicking "Reset Tree" resets an adjusted tree display back to default; clicking the "Download Tree" displays link to tree files in different formats; and clicking the "See alignment" link displays the multiple sequence alignment, where the alignment can also be downloaded via a link at the top (not shown). The results indicate that the three sample are Human Rhinovirus C, closely related to previously reported isolates named as 7384-MY-10 and 075 (G).

Human rhinovirus C isolate TN.82.11.2 Pico\_F

Human rhinovirus C isolate P114 polyprotein gene, partial cds

Human rhinovirus C isolate 03-010 capsid protein gene, partial cds

MOLE-BLAST Page 3

## **Example Searches (cont.)**

The second example concerns identification of organisms present in wastewater samples, through 16S RNA sequences.

Input sequences are uploaded through "Choose File" button and curated 16S RNA database is selected with other settings left at default (not shown).

This tree is based on MUSCLE multiple alignment computed for Mole-BLAST.

This tree is based on MUSCLE multiple alignment computed for Mole-BLAST Reset Tree ▶ Download Tree See alignment Phylogenetic Tree View D C Moleblast RID HP65KKN7413 Database nr In the resulting tree page (right), se-Tree method Collapse Mode Max Seq Difference Sequence Label Locus quence titles are highlighted by 0.75 ▼ 🔞 Sequence Title (if avail 🔻 🍪 Show All 🔻 🔞 Locus 1 🔻 🤢 Fast Minimum Evolution ▼ background color to indicate if they Mouse over an internal nod are from type culture, with colored Falstrhodobacter halotolerans strain JA744 16S ribosomal RNA gene, partial sequence Locus 2 ▼ MI17950:310-1666 Wastewater metagenome whole genome shotgun sequence Rhodobacter blasticus strain KC2138 16S ribosomal RNA gene, partial sequence dot to indicate their major taxonomic Locus 1 Rhodobacter blasticus strain ATCC 33485 16S ribosomal RNA gene, complete sequence aematobacter massiliensis strain KC2145 16S ribosomal RNA gene, partial sequence group assignment (bacterial class), Rhodobacter capsulatus strain KC2137 16S ribosomal RNA gene, partial sequence

Thiofaba tepidiphila strain BDA453 16S ribosomal RNA gene, partial sequence

Ectothiorhodospira mobilis strain DSM 237 16S ribosomal RNA gene, partial sequence

Thioalkalivibrio sulfidophilus strain HL-EbGR7 16S ribosomal RNA gene, complete s Locus 3 respectively. Color keys are listed in the Legend (A) shown to the right of PMI85:211679-212697 Wastewater metagenome whole genome shotgun sequence
Plasticicumulans lactativorans strain YD 16S ribosomal RNA gene, partial sequence
Plasticicumulans acidivorans strain TUD-Y337 16S ribosomal RNA gene, partial sequence the tree. The displayed group "Locus 1" contains eight of the eleven input from type material Pedobacter rhizosphaerae strain 01-96 16S ribosomal RNA gene, Pedobacter roseus strain CL-GP80 16S ribosomal RNA gene, Pedobacter terrae strain DS-57 16S ribosomal RNA gene, partis Pedobacter suwonensis strain 15-52 16S ribosomal RNA gene, queries (B). Results for other queries Blast names color map not displayed initially are available a-proteobacteria through the Locus pull down menu Mucilaginibacter gynuensis strain YC7004 16S ribosomal RNA

Cytophaga fermentans strain NBRC 15936 16S ribo В (C). Clicking the "See alignment" (D) Marinifilum flexuosum strain M30 16S ribosomal R Mariniflum flexuosum strain M.30 10s ribosomal R
Mariniflum flegile strain IC2469 16s ribosomal R
APMI23720:407-1444 Wastewater metagenome whole gen
Weenweeksia hongkongensis strain DSM 17368 16
Owenweeksia hongkongensis strain UST0020801
Flectobacillus major strain ATCC 29496 16S ribosomal RNA gene, partial sequence and the strain UST040 flower of the strain SWLS 16S ribosomal RNA gene, partial sequence and provided the strain SWLS 16S ribosomal RNA gene, partial sequence g-proteobacteria link retrieves the underlying multiple CFB group bacteria alignment for the tree (E). The unit of high GC Gram+ the tree (F) represents expected Arcicella rigui strain NSW-5 16S ribosomal RNA gene, partial sequence
 Arcicella aurantiaca strain TNR-18 16S ribosomal RNA gene, partial sequence number of changes per 100 bases. Flectobacillus lacus strain CL-GP79 16S ribosomal RNA gene, partial seque The tree is rooted at the middle of | Decuporation size as statistical statistics and the statistics and t the longest edge. Runella defluvii strain EMB13 16S ribosomal RNA gene, partial sequence

APMI56016:1359-1740 Wastewater metagenome whole genome shotgun sequence

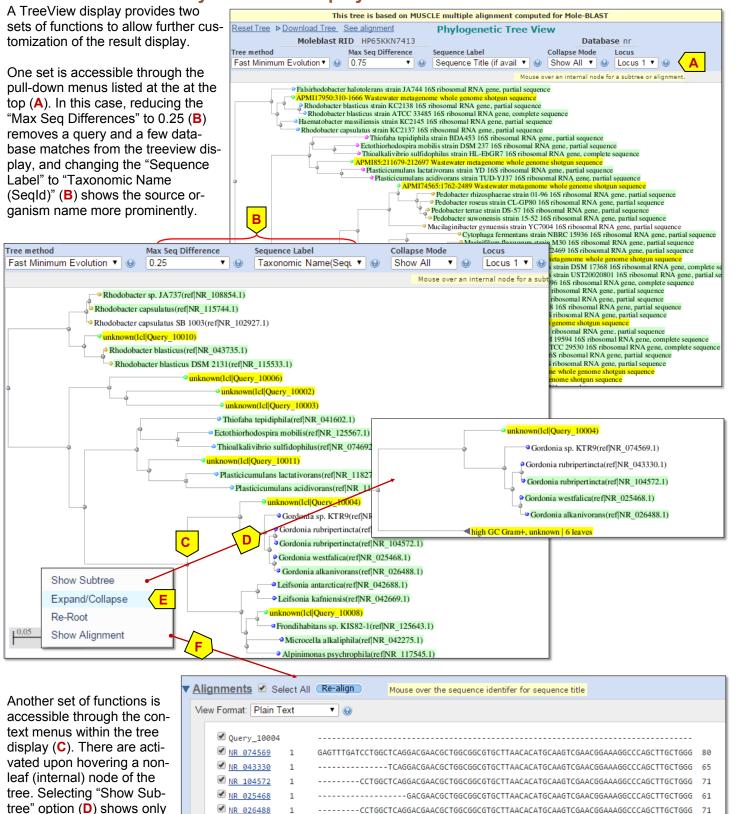
APMI5559:1-997 Wastewater metagenome whole genome shotgun sequence

Gordonia sp. KTR9 strain KTR9 16S ribosomal RNA, complete sequence Download alignment Fasta plus gaps Clustal Phylip Nexus ASN.1 Gordonia sp. K1 RV strain K1 RV 165 ribosomal RNA, compiete sequence
Gordonia rubriperincta strain DSM 4324 165 ribosomal RNA gene, partial sequence
Gordonia westfalica strain Kb2 165 ribosomal RNA gene, complete sequence
Gordonia westfalica strain Kb2 165 ribosomal RNA gene, partial sequence
Gordonia alkanivorans strain HKI 0136 165 ribosomal RNA gene, partial sequence
APMI29102:16-1530 Wastewater metagenome whole genome shotgum sequence
Galbitalea soli strain KIS82-1 165 ribosomal RNA gene, partial sequence
Altinipropara genebosobile strain CG 35 165 ribosomal RNA gene, partial sequence E Back to tree ▶ <u>Download</u> MoleBlast RID: KWeC4cr2qw7gX Alpinimonas psychrophila strain Cr8-25 16S ribosomal RNA gene, partial sequence
 Klugiella xanthotipulae strain 44C3 16S ribosomal RNA gene, partial sequence
 Leifsonia antarctica strain SPC-20 16S ribosomal RNA gene, complete sequence ▼ Descriptions Links Accession Microcella alkaliphila strain AC4r 16S ribosomal RNA gene, complete sequence In the alignment view, the ✓ NR 042275.1 ✓ NR 042688.1 Leifsonia antarctica strain SPC-20 16S ribosomal RNA gene, complete sequence Descriptions section lists ✓ NR 042669.1 the sequence titles, and the Leifsonia kafniensis strain KFC-22 16S ribosomal RNA gene, complete sequence ✓ NR 117545.1 Alpinimonas psychrophila strain Cr8-25 16S ribosomal RNA gene, partial sequence Alignments section shows ✓ Icl|Query\_10008 APMI29102:16-1530 Wastewater metagenome whole genome shotgun seguence the details of aligned se-Galbitalea soli strain KIS82-1 16S ribosomal RNA gene, partial sequence ✓ NR 125643.1 quences (both shown in ✓ Icl|Query\_10004 APMI53559:1-997 Wastewater metagenome whole genome shotgun sequence part to the left). The align-✓ NR 074569.1 Gordonia sp. KTR9 strain KTR9 16S ribosomal RNA, complete sequence ment can be redone using ✓ NR 043330.1 Gordonia rubripertincta strain DSM 43248 16S ribosomal RNA gene, partial seguence the "Re-align" button (G), ✓ NR 104572.1 Gordonia rubripertincta strain N4 16S ribosomal RNA gene, complete sequence after deselecting certain NR 025468.1 Gordonia westfalica strain Kb2 16S ribosomal RNA gene, partial sequence undesired entries using their checkboxes. The alignment file can be down-🔻 Alignments 🗹 Select All 🛛 Re-align Mouse over the sequence identifer for sequence title loaded by clicking the View Format: Plain Text "Download" link (H), which opens a section listing ✓ NR 042275 -----TGATCCTGGCTCAGGACGACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGATGAA--C available formats. ✓ NR 042688 -----AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGATGAA---✓ NR 042669 ---TTTGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGATGAA---✓ NR 117545 ---AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGATGAA--G ☑ Query\_10008 ----AGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGATGAAAGC 69 ✓ NR 125643 -----ATGGCTCAGGATGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGGTGAACGA Ouery\_10004 ✓ NR 074569 --GAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGGAAAG---✓ NR 043330 ------TCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGGAAAG---✓ NR 104572 -----CCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGGAAAG---✓ NR 025468 -----GACGAACGCTGGCGGCGTGCTTAACACATGCAA-GTCGAACGGAAAG---

Contact: blast-help@ncbi.nlm.nih.gov

Page 4 MOLE-BLAST

## **Functions Provided by TreeView Display**



a zoomed-in view. Selecting the "Expand/Collapse" option (**E**) collapses the branch under that node. Selecting the "Show Alignment" option (**F**) changes the display to show the alignment for that branch.

## **Technical Assistance**

that selected tree branch in

Please send questions and feedback on MOLE-BLAST to the blast-help group at: blast-help@ncbi.nlm.nih.gov